Ethics and Information Technology (2005) 7:149–155
DOI 10.1007/s10676-006-0004-4

Artificial morality: Top-down, bottom-up, and hybrid approaches

Colin Allen[1], Iva Smit[2] and Wendell Wallach[3]

[1]*Department of History and Philosophy of Science, 1011 E. Third Street, Bloomington, IN 47405, USA*
[2]*E&E Consultants, Slawijkseweg 11, 7077 AM, Netterden, The Netherlands*
[3]*Yale University Interdisciplinary Center for Bioethics, P.O. Box 208209, New Haven, CT 06520-8209, USA*
E-mails: colallen@indiana.edu; iva.smit@planet.nl; wwallach@comcast.net

Abstract. A principal goal of the discipline of artificial morality is to design artificial agents to act as if they are moral agents. Intermediate goals of artificial morality are directed at building into AI systems sensitivity to the values, ethics, and legality of activities. The development of an effective foundation for the field of artificial morality involves exploring the technological and philosophical issues involved in making computers into explicit moral reasoners. The goal of this paper is to discuss strategies for implementing artificial morality and the differing criteria for success that are appropriate to different strategies.

Key words: artificial morality, autonomous agents, ethics, machines, robots, values

Introduction

Artificial morality shifts some of the burden for ethical behavior away from designers and users, and onto the computer systems themselves. The task of developing artificial moral agents becomes particularly important as computers are being designed to perform with greater and greater autonomy, i.e., with less and less direct human supervision. The speed at which computers execute tasks increasingly prohibits humans from evaluating whether each action is performed in a responsible or ethical manner. Implementing software agents with moral decision-making capabilities offers the promise of computer systems that are able to evaluate whether each action performed is ethically appropriate. This in no way serves as a substitute for the moral responsibility of those who deploy or use computers. It merely means that it will be easier to use computers in an ethical manner if sensitivity to ethical and legal values is integral to the software (Wallach 2004).

Regardless of whether artificial morality is genuine morality, artificial agents act in ways that have moral consequences. This is not simply to say that they may cause harm – even falling trees do that. Rather, it is to draw attention to the fact that the harms caused by artificial agents may be monitored and regulated by the agents themselves. While some aspects of the insertion of intelligent artifacts into the moral landscape may be manageable without specific attention to the software or mechanisms controlling their behavior – i.e., by treating them as ethical 'black boxes' – we would argue that artificial morality should also be approached proactively, as an engineering design challenge for explicitly building ethically appropriate behavior into artificial agents. The attempt to design artificial moral agents forces consideration of the information required for ethical decision making, and of the algorithms which may be appropriately applied to the available information.

In this paper we discuss the philosophical roots and computational possibilities of top-down and bottom-up strategies for designing artificial moral agents (AMAs). We pick up where Allen et al. (2000) left off when they wrote, 'Essential to building a morally praiseworthy agent is the task of giving it enough intelligence to assess the effects of its actions on sentient beings, and to use those assessments to make appropriate choices'. Top-down approaches to this task involve turning explicit theories of moral behavior into algorithms. Bottom-up approaches involve attempts to train or evolve agents whose behavior emulates morally praiseworthy human behavior.

Top-down approaches

The idea behind top-down approaches to the design of AMAs is that moral principles or theories may be used as rules for the selection of ethically appropriate actions. Rule-based approaches to artificial

intelligence have been appropriately criticized for their unsuitability for providing a general theory of intelligent action. Such approaches have proven to be insufficiently robust for almost any real-world task. Yet there remain specific domains where rule-based approaches provide the best available technology, and it is an open research question whether moral behavior is one of these domains. Hence, even though (we believe) the objections to rule-based approaches are rather strong, it is incumbent on aspiring designers of AMAs to consider the prospects and problems inherent in top-down approaches. Specific objections to top-down approaches to artificial morality (rather than general objections to rule-based A.I. generally) are best understood after giving careful consideration to the prospects for building AMAs by implementing decision procedures that are modeled on explicit moral theories.

Candidate principles for conversion to algorithmic decision procedures range from religious ideals and moral codes to culturally endorsed values and philosophical systems. The Golden Rule, The Ten Commandments, utilitarianism, and Kantian deontology are some of the possible sources of rules for top-down ethical systems. When considering robots, Asimov's three laws (which he later expanded to four) also come to mind as a top-down system of ethics. While many of the same values are evident in differing ethical systems, there are also significant differences which would make the selection of any particular theory for top-down implementation a matter of obvious controversy. We will not enter that controversy here – instead, our task is to be equal-opportunity critics, pointing out strengths and limitations of some major top-down approaches.

Abstractly considered, top-down morality is all about having a set of rules to follow. In some ways of thinking, the list of rules is a heterogeneous grab bag of whatever needs to be specifically proscribed or prescribed. This is the 'commandment' model of morality, which, as well as having roots in the Judaic tradition, also pops up in Asimov's laws of robotics. A strength of commandment models is that they can have particular rules tailored to particular types of ethically relevant behavior (one for killing, one for stealing, etc.). However, a major trouble with commandment models, thus conceived, is that the rules often conflict. Such conflicts produce computationally intractable situations unless there is some further principle or rule for resolving the conflict. Most commandment systems are silent on how conflicts are to be resolved. Asimov's initial approach was to prioritize the rules so that the first law always trumped the second, which in turn always trumped the third. Unfortunately, however, the first two of Asimov's original laws are each sufficient to produce intractable conflict on their own. The addition of the 'zeroth' law, to protect humanity as a whole, is unfortunately silent on what counts as such a harm, and thus does not effectively arbitrate when there are mutually incompatible duties to prevent harm to different individuals, i.e., when protecting one will cause harm to the other and vice versa.

Philosophers have attempted to find more general or abstract principles from which all the more specific or particular principles might be derived. Both utilitarianism and Kantian deontology provide general principles which may be relevant to the design of ethical robots (Gips 1995).

A strength of utilitarianism in a computational context is its explicit commitment to quantifying goods and harms. This is also, of course, a weakness – for it is a notorious problem of utilitarianism that certain pleasures and pains appear to be incommensurable. While some economists may think that money provides a common measure (how much one is willing to spend to obtain some good or avoid some harm), this is controversial. But even if the problem of measurement could be solved, any top-down implementation of utilitarianism would have a lot of computing to do. This is because many, if not all, of the consequences of the available alternatives must be computed in advance in order to compare them. Consequences of acts will range over varying types of members of the moral constituency (people, perhaps some animals, and possibly even entire ecosystems), and many secondary 'ripple' effects will have to be anticipated. The utilitarian AMA may also have to decide whether and how to discount effects in the distant future.

In contrast to utilitarianism, deontological theories focus on the motives for action, and require agents to respect specific duties and rights. To resolve the problem that specific duties may appear to conflict, all prima facie duties may be submitted to a higher principle, such as Kant's categorical imperative. It turns out that a computational Kantian would also have to do a lot of computing in order to achieve a full moral evaluation of any action. This is because Kant's approach to the moral evaluation of actions requires not only access to one's motives, which an AMA might not have, but also a full understanding of how to characterize the motives behind the action, and an assessment of whether there would be any inconsistency if every rational agent, including humans, acted on the same motive. This requires a lot of understanding of human psychology and of the effects of actions in the world. The problem is not unique to Kant's theory – other general deontological

principles that seek to resolve conflicts among prima facie duties would face similar issues.

Both consequentialist (e.g., utilitarian) and deontological (e.g., Kantian) approaches raise their own specific computational problems, but they also raise a common problem of whether any computer (or human, for that matter) could ever gather and compare all the information that would be necessary for the theories to be applied in real time. This problem seems especially acute for a consequentialist approach, since the consequences of any action are essentially unbounded in space or time. The problem does not go away for a deontologist because consistency between the duties can typically only be assessed through their effects in space and time.

Of course humans apply consequentialist and deontological reasoning to practical problems without calculating endlessly the utility or moral ramifications of an act in all possible situations. Our morality, just as our reasoning, is bounded by time, capacity, and inclination. In a similar vein, parameters might also be set on the extent to which a computational system analyzes the consequences or imperative of a specific action. How might we set those limits on the options considered by a computer system, and will the course of action taken by such a system in addressing a specific challenge be satisfactory? In humans the limits on reflection are set by heuristics and affective controls. Both heuristics and affect can at times be irrational, but also tend to embody the wisdom gained through experience. We may well be able to implement heuristics in computational systems. Professional codes of conduct may be of some help in this context (e.g., Floridi and Sanders 2004 argue that Association of Computing Machinery Code of Ethics may be adapted for artificial agents). Nevertheless, heuristic rules of thumb leave many issues of priority and consistency unresolved. The implementation of affective controls represents a much more difficult challenge.

Bottom-up and developmental approaches

By 'bottom-up' approaches to the development of AMAs we mean those that do not impose a specific moral theory, but which seek to provide environments in which appropriate behavior is selected or rewarded. These approaches to the development of moral sensibility entail piecemeal learning through experience, either by unconscious mechanistic trial and failure of evolution, the tinkering of programmers or engineers as they encounter new challenges, or the educational development of a learning machine. Each of these methods shares some characteristics with the manner in which a young child acquires a moral education in a social context which identifies appropriate and inappropriate behavior without necessarily providing an explicit theory of what counts as such.

Bottom-up strategies hold the promise of giving rise to skills and standards that are integral to the over-all design of the system, but they are extremely difficult to evolve or develop. Evolution and learning are filled with trial and error – learning from mistakes and unsuccessful strategies. This can be a slow task, even in the accelerated world of computer processing and evolutionary algorithms.

Alan Turing reasoned in his classic paper 'Computing machinery and intelligence' (Turing 1950) that if we could put a computer through an educational regime comparable to the education a child receives, 'We may hope that machines will eventually compete with men in all purely intellectual fields'. Presumably this educational regime might include a moral education similar to the manner in which we humans acquire a sensibility regarding the moral ramifications of our actions. But simulating a child's mind is only one of the strategies being pursued for designing intelligent agents capable of learning. Simulations of evolution, the design of computational learning platforms, and associative learning techniques will all play a role in the bottom-up development of AMAs. We'll begin with a discussion of the more primitive approaches, for they might contribute to the design of the more sophisticated learning platforms.

Artificial life and the emergence of social values

The platform for artificial life experiments (Alife) – i.e., the simulation of evolution within computer systems – is the unfolding of a genetic algorithm within relatively simply defined computer environments. An intriguing question is whether a computer or Alife might be able to evolve ethical behavior. The prospect that Alife could give birth to moral agents derives from Wilson's (1975) hypothesis that a science of sociobiology might give rise to 'a precise account of the evolutionary origin of ethics'. If the foundational values of human society are rooted in our biological heritage, than it would be reasonable to presume that these values could reemerge in a simulation of natural selection.

The most promising Alife experiments derive from game theory and involve iterated rounds of the prisoners' dilemma game (Axelrod and Hamilton 1981). Danielson (1992, 1998) and his colleagues at the University of British Columbia's Centre for Applied Ethics constructed simulated environments in which virtual organisms can change and adapt in response

to the actions of other entities in the population. Danielson calls his Alife simulations 'moral ecologies'. As in the prisoners' dilemma, these organisms can cooperate or defect, act as predators or altruists. They move about within the computer simulation, and to Danielson's surprise the mindless individual entities began to form their own groups. The altruists would group together with other altruists and the predators would also hang out together. In tough times, when the resources were limited, the predators would die off while the cooperators had a competitive advantage. Danielson proposes a concept of 'functional' morality in which rationality is the only prerequisite for an agent to be a moral agent (see also Skyrms 1996, and Danielson and Harms' Evolving Artificial Moral Ecologies project, accessed at http://www.ethics.ubc.ca/eame/ on September 17, 2005).

Alife simulations hold great promise in fostering the emergence of moral agents that display at least foundational values in their behavior. However, to date, the mindless moral agents developed within simulations of evolution are very simple and far from being able to engage in reflection on abstract and theoretical issues, which is a hallmark of the sophisticated moral sensibilities of humans. The question is still outstanding whether Alife will prove to be helpful in the development of artificial moral agents capable of engaging the more complex dilemmas that we encounter daily. We should not forget, however, that evolution has also led to immoral behavior.

Unbiased learning platforms

Noting the limitations inherent in systems designed around rule-based ethics, Chris Lang[1] recommends 'quest ethics', a strategy wherein the computer learns about ethics through a never-ending quest to maximize its goal, whether that goal is to be 'just' or to be 'moral'. By endlessly searching for a better and better solution, such a machine will presumably develop from a temporarily immature state to a level where we might designate it a moral agent.

Lang's 'unbiased learning machines' are designed around a non-terminating learning algorithm or what is sometimes called a 'hill-climbing' or a 'greedy-search' algorithm. In straightforward applications, these algorithms evolve to be smarter and demonstrably faster than their human counterparts. Lang is optimistic, but can a machine, whose actions are determined by its previous state and inputs, gain the freedom to choose between options, to model itself to an environment, and act in accordance with human values?

There are two central challenges in designing a computer capable of continually questing for a higher morality: specifying the goal or goals of the system, and enabling an endless flow of fresh real world data that expand the domain the system peruses in its quest. Although defining the goals could lead to considerable philosophical disagreement, the more difficult challenge in designing a morally praiseworthy agent lies in stimulating the system to expand its realm of potential choices. For as von Foerster (1992) pointed out, it is not just the question of being able to choose, but also the expansion of the available choices that is central to our ethics and our human ability to invent ourselves.

Associative learning machines and humanoids

If morality is primarily learned, built up through experience, trial and error, and honed through our capacity for reason, then teaching an AMA to be a moral agent may well require a similar process of education. Associative learning techniques model the education children receive in the form of rewards and punishment, approval and disapproval. Finding effective counterparts to reward and punish, suitable for training a robot or computer is a problematic task. Lang recommends rewarding unbiased learning machines with richer data when they behave ethically. Simulating sensory pain is also an option. The intent behind a reward and punishment system of moral education may be that the child's discovery of shared ethical concerns or principles may be most effectively guided by a process of discovery that is guided by parents and teachers who have at least a rudimentary understanding of the trajectory the learning will follow. Children naturally move on to the next level of moral reasoning as they come to appreciate the limitations of the reasons which they have identified. The primary challenge of imbuing an AMA with such a degree of insight lies in implementing feedback to the computation system that goes beyond a simple binary indicator that a justification for an act is acceptable or unacceptable.

In a controlled laboratory we would need to create a series of learning situations through which the AMA would work its way toward a level of moral behavior acceptable by the standards we define. The rich plethora of algorithms developed to facilitate machine learning, case-based reasoning, data acquisition, and data-mining are just beginning to be combined into computational systems capable of utilizing an array of learning tools. In principle there

[1] C. Lang, Ethics for Artificial Intelligences, unpublished ms. accessed at http://philosophy.wisc.edu/lang/AIEthics/index.htm on September 17, 2005.

is no reason why these learning platforms cannot be adapted for at least rudimentary moral reasoning. If we hope to pursue a developmental path in creating an AMA, whether our artificial agent actually learns in the same manner as we humans is less important than their ability to learn. The more serious consideration is whether machines can be trained to work with the kind of abstract principles that are the hallmark of higher order moral reasoning.

Dangers inherent in learning systems

Chris Lang's optimistic vision that unbiased learning systems would develop naturally toward an ethical sensibility that valued humans and human ethical concerns sits in sharp contrast to the more dire futuristic predictions regarding the dangers AI poses. After all, any system which has the ability to learn can also potentially undo any restraints built into the system. The layered architecture of computational systems commonly isolates lower level standards and protocols from higher order functionality. Core restraints can presumably be built into foundation layers of the computer platform, which are inaccessible to those parts of the computer that learn and revise the structures that filter new information. This of course raises the question of what moral restraints or moral grammar are to be encoded into these 'deeper' protocols.

The concept of a moral grammar plays a key role in Josh Storrs Hall's discussion of machine ethics.[2] In his view, moral codes are similar to language grammars, for the vast majority of people 'learn moral rules by osmosis, internalizing them not unlike the rules of grammar of their native language, structuring every act as unconsciously as our inbuilt grammar structures our sentences'. He argues that there are structures in our brains that predispose us to learn moral codes, and that determine within broad limits the kinds of codes we can learn. Thus 'while the moral codes of human cultures vary within those limits, they have many structural features in common'. Extensive studies notwithstanding, the innate grammar of language has remained elusive, as, can be argued, has the 'moral deep structure', or what the philosophers of the Scottish Enlightenment called the 'moral sense'. It would require a clear explication of the code that governs our moral sensibility to program this foundational grammar into the lower levels of a system's architecture.

[2] J. Storrs Hall, Ethics for Machines, unpublished ms. Accessed http://discuss.foresight.org/~josh/ethics.html on September 17, 2005.

Key restraints could also be programmed into a computational system at a very low level. They would act as something like a human conscience. In the short-term we need not be concerned that a learning system will root out these deeply embedded restraints. But like the ability of humans to override their conscience given the right goals, desire, or motivation, learning computers might also find ways to circumvent restraints that got in the way of their goals. In the meantime, the larger concern is that very small incremental changes made to structures through which an AMA acquires and processes the information it is assimilating will lead to subtle, disturbing, and potentially destructive behavior. In situations requiring precision, even little errors could potentially lead to dramatic consequences. Learning systems may well be one of the better options for developing sophisticated AMAs, but the approach holds its own set of unique issues.

Hybrid approaches

The top-down approaches emphasize the importance of explicit ethical concerns that arise from outside of the entity, while the bottom-up approaches are directed more at the cultivation of implicit values that arise from within the entity. Top-down principles represent broad controls, while values that emerge through the bottom-up development of a system can be understood as causal determinants of a system's behavior. Ethical principles such as justice and maximizing the aggregate good tend to restrict options – they presume a context in which the actor has broad freedoms in the manner he can act, but whose action must be confined to morally praiseworthy behavior. Evolution and the learning of a machine are directed toward the expansion of choices and flexibility in behavior. The ethical restraints the evolving system learns to honor are those that will increase its choices and its opportunity to survive and flourish. The top-down ethical restraints reinforce the principle that moral behavior often requires limiting one's freedom of action and behaving in ways that may not be in one's short-term or self-centered interest for the good of society. Both top-down and bottom-up approaches embody different aspects of what we commonly consider a sophisticated moral sensibility.

If no single approach meets the criteria for designating an artificial entity as a moral agent, then some hybrid will be necessary. Hybrid approaches pose the additional problems of meshing both diverse philosophies and dissimilar architectures. Genetically acquired propensities, the rediscovery of core values through experience, and the learning of culturally

endorsed rules all influence the moral development of a child. During young adulthood those rules may be reformulated into abstract principles that guide one's behavior. We should not be surprised if designing a praiseworthy moral agent will also require computational systems capable of integrating diverse inputs and influences, including top-down values informed by a foundation moral grammar and a rich appreciation of context. Von Neumann Machines and neural networks, genetic and learning algorithms, rule and natural language parsers, virtual machines and embodied robots, affective computing and standards for facilitating cooperative behavior between dissimilar systems may all be enlisted in tackling aspects of the challenge entailed in designing an effective AMA. Designers of AMAs cannot afford to be theoretical purists with respect to questions about how to approach moral intelligence.

One central question is whether systems capable of making moral-decisions will require some form of emotions, consciousness, a theory of mind, an understanding of the semantic content of symbols, or need to be embodied in the world. While we feel that hybrid systems without affective or advanced cognitive faculties will be functional in many domains, it will be essential to recognize when additional capabilities will be needed. Embodiment is both a goal and a bottom-up strategy that is very evident in subsumptive and evolutionary or epigenetic robotic architecture. The promise that other advanced affective and cognitive skills will emerge through the evolution of complex embodied systems is highly speculative. Nevertheless, facilitating the emergence of advanced faculties or designing modules for complex affective and cognitive skills may well be required for fully autonomous AMAs.

Evaluating machine morality

As we stated above, the goal of the discipline of artificial morality is to design artificial agents to act as if they are moral agents. Questions remain about appropriate criteria for evaluating effectiveness in this area. Just as there is not one universal ethical theory, there is no agreement on what it means to be a moral agent, let alone a successful artificial moral agent. A Moral Turing Test (Allen et al. 2000) is one possible strategy for evaluating the adequacy of an AMA in light of differing theories of moral agency. Turing's test for machine intelligence is notoriously controversial, and we would not endorse it as a criterion for strong A.I. or genuine moral agency. Nevertheless, some sort of 'blind' comparison to human performance may be a useful tool in assessing the acceptability of AMA behavior.

It is important to keep in mind that the existing ethical theories were developed with human beings in mind, long before computers and autonomous intelligent agents entered the scene. The engineering requirements of artificial morality place the discordant theories in a new context that forces us to re-evaluate their significance as specifications for moral action. Computers and robots are largely a product of a materialistic worldview which presupposes a set of metaphysical assumptions that are not always compatible with the spiritual worldviews which produced many of our ethical categories and much of our ethical understanding. While this metaphysical tension by no means obviates the project of introducing moral decision-making abilities into computers, it should sensitize us about being too facile in reducing complex subjects, such as the need for a conscience, to an easily manageable set of skills. Honest people may differ on the extent to which human morality may be transplanted to machines. But the challenges posed by our increasingly autonomous systems will not go away, even if they mean that a new ethical theory has to be developed for artificial morality. This is a topic which requires further discussion.

Computers pose unique ethical considerations that are of particular philosophical interest (Floridi 1999). The field of artificial morality extends that interest beyond mere theory and analysis and towards an actual engineering task. The objective of building morally praiseworthy systems is complementary to that of distributing and enforcing accountability when things go wrong. However, possibilities for 'creative stewardship' of complex artifacts (Floridi and Sanders 2004) will be enhanced if, rather than viewing artificial moral agents as ethical black boxes, engineers, programmers, and philosophers collaboratively contribute to the analysis of the engineering requirements for the development of artificial morality.

References

C. Allen, G. Varner and J. Zinser. A Prolegomena to Any Future Artificial Moral Agent. *Journal of Experimental and Theoretical Artificial Intelligence*, 12: 251–261, 2000.

R. Axelrod and W. Hamilton. The Evolution of Cooperation. *Science*, 211: 1390–1396, 1981.

P. Danielson, *Artificial Morality: Virtuous Robots for Virtual Games*. Routledge, New York, 1992.

P. Danielson, *Modeling Rationality, Morality and Evolution*. Oxford University Press, New York, 1998.

L. Floridi. Information Ethics: On the Philosophical Foundation of Computer Ethics. *Ethics and Information Technology*, 1: 37–56, 1999.

L. Floridi and J.W. Sanders. On the Morality of Artificial Agents. *Minds and Machines*, 14: 349–379, 2004.

H. Foerster. Ethics and Second-order Cybernetics. *Cybernetics & Human Knowing*, 1: 9–25, 1992.

J. Gips. Towards the Ethical Robot. In K. Ford, C. Glymour and P. Hayes, editors, Android Epistemology, pp. 243–252. MIT Press, Cambridge, MA, 1995.

B. Skyrms, *Evolution of the Social Contract*. Cambridge University Press, New York, 1996.

A. Turing. Computing Machinery and Intelligence. *Mind*, 59: 433–460, 1950.

W. Wallach. Artificial Morality: Bounded Rationality, Bounded Morality and Emotions. In I. Smit, G. Lasker and W. Wallach, editors, *Proceedings of the Intersymp 2004 Workshop on Cognitive, Emotive and Ethical Aspects of Decision Making in Humans and in Artificial Intelligence*, pp. 1–6, Baden-Baden, Germany, IIAS, Windsor, Ontario, 2004.

E.O. Wilson, *Sociobiology: The New Synthesis*. Harvard/Belknap, Cambridge, MA, 1975.